Advanced Assay Development Guidelines for Image-Based High Content Screening and Analysis

Mark-Anthony Bray, Ph.D.
Imaging Platform, Broad Institute
of MIT and Harvard
mbray@broadinstitute.org

Anne Carpenter, Ph.D.
Imaging Platform, Broad Institute
of MIT and Harvard
anne@broadinstitute.org

Imaging Platform, Broad Institute of MIT and Harvard

Abstract

Automated microscopes are now widespread in biological research. They enable the unprecedented collection of images at a rate which outpaces researchers' ability to visually inspect them. Whether interrogating hundreds of thousands of individual fixed samples or fewer samples collected over time, automated image analysis has become necessary to identify interesting samples and extract quantitative information by microscopy. This chapter builds on the material presented in the introductory HCS section.

---

1 Experimental design for HCS

1.1 Controls

Whenever possible, positive and negative controls should be included in an assay. Using controls is required to calculate a performance envelope for measuring the assay quality and phenotype feature space (see "Assay Quality and Acceptance Criteria for HCS" section below).

However, positive controls may not be readily available. In these situations, an assay measuring a real biological process may still show a phenotype of interest under some conditions that can be observed and measured even if positive controls that induce high levels of cells with the phenotype do not exist. Once a condition is identified and demonstrates such a measurable change, then it can serve as a positive control going forward.

For profiling assays, in which a large variety of cellular features are measured to identify similarities among samples, and hence designed to have multiple readouts, several different positive controls for each desired class of outcomes may be necessary. However, these may not be known in advance. Long running assays will typically accumulate positive controls over time and may even change the perceived limits or dynamic range of the assay.

Ideally, a positive control is of the same type as the reagents to be screened (e.g. a small molecule control for a small molecule screen, and an RNAi-based control for an RNAi screen). However, any reagent that induces the phenotypic change of interest can serve as a

positive control if necessary, even if artificial. For example, expression of a constitutively active form of a tagged protein or knockdown of a target by RNAi can simulate the effects of a sought-after small molecule in a screen. Although differences in modality may complicate direct quantitative analysis of such controls, such 'artificial' controls are often helpful during assay development and optimization, and provide a sense of the dynamic range to be expected in the screen.

In selecting controls, there is a temptation to select reagents with very strong effects as positive controls. This is often the result of undue emphasis on minimum criteria for acceptance for screening, such as a Z'-factor cutoff. Good judgment should instead prevail, and positive controls should be selected based on the intensity of the hits hoped to find. For example, it is not helpful to select a very strong positive control that yields a high-quality Z'-factor if it is not comparable to the strength of the expected hits sought in an actual screen. Instead, inclusion of moderate to mild positive controls, or decreasing doses of a strong positive control, is better in gaining a sense of the sensitivity of the assay to realistic hits.

The authors have observed several successful screens with sub-par Z'-factors or absent a positive control that nonetheless yielded high-value, reproducible, biologically relevant hits. As such, common sense should prevail by factoring in the complexity and value of hits in the screen and the degree of tolerance for false positives that can be filtered out in confirmation screens.

For plates of reagents from vendors, typically only the first and the last columns of a multi-well plate are available for controls, with the treated samples contained in the middle wells.

Unfortunately, this practice renders the assay susceptible to the well-known problem of plate-based edge effects, which lead to over- or under-estimation of cellular responses when normalizing by the control wells.

One strategy to minimize edge effects is to spatially alternate the positive and negative controls in the available wells, such that they appear in equal numbers on each of the available rows and on each of the available columns[1][2]. (Figure 1)

If the screener is creating a custom plate for an HCS run, ideally the control wells should be randomly placed across the plate in order to avoid spatial bias. However, this approach is rarely practical in large screens as it must be performed manually. Therefore, the chance of introducing a spatial bias effect by using a non-random control placement is an accepted practice due to the difficulty in creating truly random plate arrangements.

For screens run in multiple batches where the controls need to be prepared such as lentiviral shRNA screens which necessitate the creation of viral vectors, variation in viral preparation may be confounded with assay variation. One helpful strategy in this situation is to make a plate containing controls in one batch, freeze them and then thaw and use them a few at a time as the screen progresses. This method can help identify assay drift or batch specific problems.

For analytical approaches to correcting inter- and intra-plate bias, see the section "Normalization of HCS data" below.

1.2 Replicates

Like all HTS assays, the cost of replicates should be weighed against the cost of cherry-picking hits and performing a confirmation screen. HCS assays with complex phenotypes are often more difficult to score so more replicates are often needed. Experiments are normally performed in duplicate or higher replicate numbers in order to decrease both false

positive and false negative rates. Performing replicate treatments offers the following advantages[2]:

1. Taking the mean or median of the replicate measurements yields lower variability than with single measurements only.
2. Replicate measurements provide direct estimates of variability and are useful for evaluating the probability of detecting true hits. When combined with control measurements, a statistically significant change can be determined by comparing (a) the ability to distinguish treated wells from the controls, and (b) the ability to distinguish replicates of the same treatment.
3. Replicates reduce the number of false negatives without increasing the number of false positives.

Despite the multitude of good reasons for high replicate numbers, almost all large screens are performed in duplicate. Increasing the replicate number from 2 to 3 is a 50% increase in reagent cost which in the case of an HCS screen involving the tens or hundreds of thousands of samples can determine whether the screen is performed at all. HCS screens are normally performed by first screening all the samples, usually at a single concentration in duplicate, and then retesting all the hits in confirmation assays. The confirmation assays serve to filter out the false positives and are performed on a much smaller scale where it is easier to increase the replicate number and perform dose response studies if needed. In an HCS screen, preference is given to reducing false negatives because if a hit is missed during the first round of screening, it is irretrievable unless the screen is performed.

The number of treatment replicates needed is empirical and largely dictated by the subtlety of the observed cellular behavior: If a given treatment produces a strong biological response, fewer replicates will be required by virtue of a high signal-to-noise ratio. In certain cases, up to 7 replicates may be needed[3] but 2 - 4 is more typical.

Placement of replicate wells is subject to the same considerations in placement of control wells (see "Controls" above). Although randomization of the sample placement from one replicate to another is ideal, this is rarely done and for practical reasons, plates follow the same layout for all replicates. Where possible, both inter- and intra-plate replicate wells should be used for the purposes of ensuring robust normalization (see "Normalization of HCS data" below).

2 Assay Quality and Acceptance Criteria for HCS

Z'-factor (or **Z-factor**, not to be confused with z-score): While there are a number of different measurements of assay performance, the most widely used measurement to assess an assay is the so-called Z'-factor.

- *Definition*: This criteria for primary screens has gained wide acceptance in the HTS community and is defined as[4]:
  $1 - \frac{3(\sigma_p + \sigma_n)}{|\mu_p - \mu_n|}$ where μ_p and σ_p are the mean and standard deviation of the positive controls (or alternately, the treated samples) and μ_n and σ_n are those of the negative controls.
- *Range and interpretation*: The Z'-factor has the range of $-\infty$ to 1, and is traditionally interpreted as follows: (Table 1).

It should be noted that the "perfect" case of Z' = 1 implies that $\mu_p = \mu_n$ and/or $\sigma_p = \sigma_n = 0$ (e.g., no separation of the control distributions or all samples produced identical readouts in both distributions). Neither of these scenarios represent realistic assays.

For moderate assays, the screener should consider the utility of mining hits that fall into the Z' = 0 — 0.5 range. Given the screening cost of eliminating a false positive (which is hopefully low) as compared to that of eliminating a false negative (potentially very costly, as mentioned above) , a decision will need to be made whether to follow up on such hits or instead cherry-pick and repeat treatments, or re-screen entire plates.

- *Advantages*: While Z' > 0.5 has become a de facto cutoff for most HTS assays, 0 < Z' ≤ 0.5 is often acceptable for complex HCS phenotype assays, since those hits may be more subtle but still valuable. In comparison to other assay robustness metrics[5], advantages of the Z'-factor include:
  - Ease of calculation.
  - Accounts for the variability in the compared groups while properly ignoring the absolute background signal.
  - Often found in both commercial and open-source software packages.
- *Disadvantages*:
  - Does not scale linearly with signal strength. That is, an increased target reagent or a very strong positive control may achieve a higher Z'-factor that is disproportionate to the phenotype strength. The above ranges may not realistically represent more moderate screening positives which may still be biologically meaningful, e.g., RNAi screens where the signal-to-background ratio is lower than that of small-molecule screens[6].
  - The use of sample means and standard variance. Statistically, this condition assumes that the negative and positive control values follow a normal (i.e., Gaussian) distribution. The presence of outliers or asymmetry in the distributions can violate this constraint. Such is often the case for cell-based assays but is rarely verified, and can yield a misleading Z'-factor. Conversely, attempting to correct for this by transforming the response values to yield a normal distribution (e.g., log scaling) may yield an artificially high Z'-factor[7].
  - The sample mean and sample standard deviation are often not robust estimators of the distribution mean and standard deviation. In the presence of outliers, these statistics can easily lead to an inaccurate measure of control distribution separation.

One-tailed Z' factor: This measure is a variant of the Z'-factor formulation which is more robust against skewed population distributions. In such cases, long tails opposite to the mid-range point lead to a high standard deviation for either population, which results in a low Z' factor even though the population means and samples between the means may be well-separated (unpublished work).

- *Definition:* This statistic has the same formulation as the Z'-factor, with the difference that only those samples that lie between the positive/negative control population medians are used to calculate the standard deviations.
- *Range and interpretation*: Same as that for the Z'-factor.
- *Advantages*
  - Attempts to overcome the Gaussian limitation of the original Z'-factor formulation.
  - Informative for populations with moderate or high amounts of skewness.
- *Disadvantages*

  - Still subject to the scaling issues described above for the original Z'-factor formulation.
  - Not available as part of most analysis software packages.

V-factor: The V-factor is a generalization of the Z'-factor to a dose-response curve[8].

- *Definition*: Calculated as either:
  $1 - 6\frac{\sigma_{fit}}{|\sigma_p - \sigma_n|}$ where $\sigma_{fit} = \sqrt{\frac{1}{N}\sum_{i=0}^{n}(f_{exp} - f_{mod})^2}$, i.e, the root-mean-square deviation of a logistic model to the response data, and σ_p and σ_n are defined as above; or
  $1 - 6\frac{mean(\sigma)}{|\sigma_p - \sigma_n|}$ if no model is used, i.e., the average of several replicates where σ are the standard deviations of the data.
- *Range and interpretation*: Same as that for the Z'-factor.
- *Advantages*
  - Decreased susceptibility to saturation artifacts, which reduce the variability of the controls.
  - Taking the entire response curve into account makes the V-factor robust against dispensing errors (which typically occur towards the middle of the dose curve, rather than the extremes as for the positive/negative controls).
  - The V-factor formula has the same value as the Z'-factor if only two doses are considered.
- *Disadvantages*:
  - Requires dose response data, which requires many more samples than statistics relying solely on positive and negative controls.
  - Not available as part of most analysis software packages (CellProfiler is an exception).

Strictly standardized mean difference (**SSMD**, denoted as β): This measure was developed to address limitations in the Z' factor in experiments with control of moderate strength.

- *Definition*: The SSMD measures the strength of the difference between two controls, using the formulation is[9][10]:
  $\beta = \frac{\mu_n - \mu_p}{\sqrt{\sigma_n^2 + \sigma_p^2}}$ where μn, σ_n^2, μp and σ_p^2 are defined as above for the Z' factor.
- *Range and interpretation*: Acceptable screening values for SSMD depend on the strength of the positive controls used, as described in the following table[11] (these threshold values assume that the positive control response is larger than that of the negative control; if the converse is true, the threshold values are negative and the inequality signs are reversed): (Table 2).

  Zhang et al[10] make the following recommendations to choosing the appropriate criterion:

  - In chemical compound assays (which typically have positive controls with strong/extremely strong effects), use criterion (4) or (3).
  - For RNAi assays in which cell viability is the measured response, use criterion (4) for the empty control wells (i.e, wells with no cells added).
  - If the difference is not normally distributed or highly skewed, use criterion (4).
  - If only one positive control is present in the experiment, use criterion (3).
  - For two positive controls, use criterion (3) for the stronger control and criterion (2) for the weaker control.
- *Advantages*[6]
  - Ease of calculation.
  - Accounts for the variability in the compared groups.
  - Accommodates the effect size of the controls, through the use of different thresholds.
  - Lack of dependence on sample size.
  - Linked to a rigorous probability interpretation.
- *Disadvantages*[6]
  - Not available as part of most analysis software packages (CellProfiler is an exception).
  - Not intuitive for many biologists.
  - The thresholds are based on a subjective classification of control strength.

Area under the receiver operating characteristic curve (AUC):[12].

- *Definition*: The receiver characteristic curve (ROC) is a graph showing the proportion of true and false positives given a range of possible thresholds between the positive and negative control distributions (see figure). This pictorial information can be summarized as a single value by taking the area under the ROC curve (also known as the AUC) (Figure 2).
- *Range and interpretation*: The AUC can assume a value between 0 and 1. An assay which generates both true and false positives at random would result in a diagonal line between (0, 0) and (1, 1) on the ROC. For such a case, the AUC would equal 0.5. Therefore, a usable assay must therefore have an AUC > 0.5 (and ideally much higher) although no cutoff criteria has been agreed upon by the community (Table 3).

  Given that most screens will require a false positive rate of less than 1%, the right side of the ROC is typically less relevant than the left-most region. An alternate metric is to calculate the AUC from only the left-most region[12].
- *Advantages* [6]
  - Allows for the viewing the dynamic range of the data given positive and negative controls.
  - Does not assume that control distributions are normal (i.e, Gaussian)
  - Multiple thresholds for defining positives and the resulting trade-offs between true positive and true negative detection can be evaluated simultaneously.
- *Disadvantages*[6]

- Requires a large sample size for calculation. Ideally, many replicates of positive and negative controls are needed.
- Some information is lost when the AUC is used exclusively rather than visual inspection of the complete ROC. For example, two classifiers under consideration may have the same AUC but one may do better than the other at different parts of the ROC graph (that is, their curves intersect at some point). In this case, the relative accuracy no longer becomes a measure of the global performance, and restricting the AUC calculation to only a portion of the ROC graph is recommended.
- Not available as part of most analysis software packages (though GraphPad's Prism is a commonly-used exception).

3 Quality Control for HCS

Unlike most HTS assays, HCS data can be visually and sometimes automatically checked to identify and remove artifacts. In HCS assays, several fluorescent probes are often used simultaneously to stain cells, each labeling distinct cellular components in each sample. For screens in which the goal is to identify a small number of hits for a particular, known phenotype of interest, the candidate hits can often be screened by eye to eliminate false-positive artifacts. However, large screens and profiling experiments probing a broad spectrum of subtle morphological responses require more automated methods to detect and remove artifacts and systematic aberrations.

In general, the best approach to avoid imaging artifacts is to adjust the image acquisition settings to optimize the image quality (see the section "Capturing a Good Image" in the introductory HCS chapter of the AGM for more details). However, despite the screener's best efforts at acquisition and sample preparation, anomalies will still appear and end up polluting otherwise high-quality microscopy data.

Common artifacts that can confound image analysis algorithms are out-of-focus images, debris, image overexposure, and fluorophore saturation, among others. Because these anomalies affect a wide variety of intensity, morphological, and textural measurements in different ways, a single quality control (QC) metric that captures all types of artifacts, without also triggering on the unusual appearance of hits of interest, is not realistic. Instead, it is recommended to use either multiple metrics targeting the various artifacts that arise, or a supervised machine-learning approach.

3.1 Using targeted features for QC

3.1.1 Cell count as a quality control measure

Depending on the experimental context, a simple measure of quality is the calculated cell count. This metric can help identify problems with the following situations:

- Per-image object segmentation: If the screener has an idea of the typical number of cells in a given image, deviations from this range at the per-image level can be indicative of improper object segmentation. An unusually low apparent number may mean that neighboring cells are getting merged together or are absent due to cell death or incorrect cell plating, whereas a high apparent count may mean that cellular objects are being split apart incorrectly or an artifact such as compound precipitation is present.
- Per-well heterogeneities: Uneven distribution of cells have effects on cellular adhesion and morphology, and in multiwell plates, low cell counts may characterize the wells of the edge of the plate. Computation and display of per-well cell counts in

a plate layout heatmap format can reveal the presence of such systemic artifacts, which can be corrected at the sample preparation stage[14].

3.1.2 Features for detecting out-of-focus images

Despite the use of autofocusing routines on automated microscopes, out-of-focus images are a common and confounding artifact in HCS. The rate of occurrence can depend on the cell type being examined and how adherent the cells are to the bottom of the well. Two measures are particularly useful in out-of-focus image detection[15]:

- **Power log-log slope (PLLS)**: This measure evaluates the slope of the power spectral density of the pixel intensities on a log-log scale; the power spectrum density shows the strength of the spatial frequency variations as a function of frequency. It is always negative, and decreases in value as blur increases and high-frequency image components are lost.
  Typical in-focus values are in the range of -1.5 ~ -2; very negative values indicate a steep slope which means that the image is composed mostly of low spatial frequencies. It is recommended to plot the PLLS for a given channel as a histogram and examine outliers that are substantially less than -2.
  - *Advantages*: Because the PLLS of natural images is relatively invariant, this metric is useful as an unbiased estimator of focus.
  - *Disadvantages*: The presence of bright artifacts (e.g, fluorescent debris) in an otherwise in-focus image can produce artificially low PLLS values (Figure 3).
- **Textural correlation**: This measure evaluates the dependency of each pixel on the grayscale value of its neighbors, for a given spatial scale. Given the proper spatial scale, this measure shows the separation between blurry/in-focus images: as the correlation of an image increases, the blurriness of the image also increases.
  Of particular importance is the choice of spatial scale: a smaller spatial scale will pick up the blurring of smaller image features first, increasing the sensitivity. In general, it should be no larger than the diameter for a typical object (e.g, nucleus, speckle) in the channel of interest.
  - *Advantages*: Performance is generally insensitive to the amount of cell coverage.
  - *Disadvantages*: Dependence upon proper *a priori* selection of spatial scale. If the scale is too small, this metric starts reflecting the smaller in-focus image features rather than the amount of blur; too large a value, and it reflects the spatial proximity of similar cellular features.

The situation is more complicated if only a portion of the image is out-of-focus rather than the entire image (e.g., a section of a confluent cell cultures lifting off and "rolling up", causing a local change in the depth of field). In this case, the difference between such images and in-focus images will not be as distinct, but a more moderate shift in the metric values may still aid in establishing a reasonable cutoff.

3.1.3 Features for detecting images containing saturated fluorescent artifacts

Saturation artifacts are another common aberration encountered in HCS. Unusually bright regions in an image can be caused by debris contamination, aggregation of fluorescent dye, and/or inappropriately high exposure time or detector gain settings. Such regions can produce inaccurate intensity measurements and may impair cell identification even when such a region is small or not terribly bright.

The following metrics can be measured and examined to detect saturation artifacts:

- A useful measure is the percentage of the image that is occupied by saturated pixels. (Here, we define *saturation* as the maximum value in the image as opposed to the maximum bit-depth allowed by the image format.) In normal cases, only a small percentage of the image is at this value. Images with a high percentage value typically indicate either bright artifacts or saturated objects (e.g., non-artifactual dead cells). Further examination is required to determine which is the case, and whether the image is salvageable.
- The standard deviation of the pixel intensity is also useful for detection of images where a bright artifact is present but is not bright enough to cause saturation (Figure 4).

3.2 Using machine learning for QC

Machine learning provides a more intuitive way to train a computer to differentiate unusable from acceptable images on the basis of a sample set of each (the training set) and is particularly useful for detection of unforeseen aberrations. PhenoRipper is an open-source unsupervised machine-learning tool that can detect major classes of similar images and can be useful for artifact detection[16]. The advantage of taking the machine learning approach is that it does not require a priori knowledge of the important features which identify the artifact(s). Disadvantages include the time required to create the training set and the expertise to run the analysis[17]

For more on machine learning applications, see the section "Machine learning for HCS" below.

4 Normalization of HCS data

In the process of performing an HCS screen, invariably small differences between samples, including replicates, appear despite the screener's best efforts at standardization of the experimental protocol. These include both systematic (i.e., stemming from test conditions and procedure) and random errors (i.e., stemming from noise). Sources of systematic error include[18]:

1. Errors caused by reagent aging, or changes to compound concentrations due to reagent evaporation or cell decay.
2. Anomalies in liquid handling, malfunction of pipettes, robotic failures and reader effects
3. Variation in incubation time and temperature differences, temporal drift while measuring multiple wells and/or multiple plates and reader effects, and lighting or air flow present over the course of the entire screen

The combination of these effects may generate repeatable local artifacts (e.g., border, row or column effects) and global drifts recognizable as consistent trends in the per-plate measurement means/medians which result in row and/or column measurements that systematically over- or underestimate expected results[19].

In addition, the cellular population context (e.g. cell density) has a profound influence of cell behavior and may account significantly to cell-to-cell variability in response to treatment[20]; correcting for these effects involve sophisticated methods of modeling the cellular population response[21].

The overall impact of these variations depends upon the subtlety of the cellular response in question. For example, even if a positioned plate layout is used (rather than a random

layout), a strong biological response may be sufficient to overcome any plate effects that may occur by virtue of a high signal-to-noise ratio.

If this is not the case, normalization is necessary to remove these systematic variations and allow the detection of variation originating from experimental treatments. The following results are ideal:

1. The feature ranges observed across different wells with the same treatment should be similar.
2. The feature distributions of the controls (whether positive or negative) should be similar.

Examples of software packages that include a variety of plate normalization techniques include GeneData Screener (http://www.genedata.com) and Bioconductor[22].

4.1 Inter-plate normalization

For screening purposes, ideally the control wells should be present on all plates. Negative controls should be present at minimum, but preferably both positive and negative, and subject to the criteria described in the "Controls" section above, with the expectation that the cellular response in the control wells is consistently similar between all samples. However, this is rarely the case; it is not uncommon for the means and standard deviations of the collected measurements to vary from plate to plate. Hence, *inter-plate normalization* is needed to reduce variability between samples across plates.

The choice of normalization method will depend on the particular assay. Methods of inter-plate normalization include the following[6][2]:

- **Fraction or percent of controls**: The most straightforward approach is division of each sample value by the mean of the (negative or positive) control measurement of interest. This approach requires a large number of controls for sufficient estimation of the mean (see "Replicates" above), and is appropriate in instances where the data is tightly distributed and close to normal. Since information about the sample variation is not included, this measure is sensitive to outliers in the controls. A more robust version of this calculation is to substitute the median for the mean (although the sample variation is still not taken into account).
- **Normalized percent inhibition**: When a reliable positive control is available, this approach is calculated as $x'_i = \frac{H - x_i}{L - x_i} \times 100\%$ where xi is the measured value, x'i is the evaluated percentage, H is the mean of the high controls, L is the mean of low controls. However, see the caveats with the use of positive controls in "Controls" section above.
- **Fraction or percent of samples**: When a high proportion of wells are expected to produce little to no response (e.g., many RNAi studies), the mean of all samples on a plate can be used for the percent of controls formula in lieu of a negative control. A more robust version of this calculation is to substitute the median for the mean, subject to the caveats above.
  - This approach is recommended if the assay lacks good negative controls that work effectively across all plates, and can provide more accurate measures due to the larger number of samples as compared to controls while reducing the need for large numbers of controls[6].
  - A variation on this approach that is unique to HCS is normalization of nuclei intensity measures by the DNA content using the mode of the DNA

intensities, in cases where the large majority of the cells are in interphase in the cell cycle.

- However, the assumption that the reagents have no biological effect should be confirmed for the particular assay. For example, it is not recommended where most of the wells have been chosen precisely because the response is differentially expressed or the overall response level between samples is changed. The following screens would violate this assumption:
  1. Confirmation screens in which phenotype-positive reagents are evaluated on the same assay plate.
  2. Primary screens targeting structurally or functionally related genes.

- **Z-score (or standard score) and robust z-score**: The z-score transforms the measurement population distribution on each plate to a common distribution with zero mean and unit variance. The formula is $x'_i = \frac{x_i - \mu}{\sigma}$ where xi is the measured value, x'i is the normalized output value and μ and σ are the mean and standard deviations, respectively. An advantage of this approach is the incorporation of the sample variation into the calculation. However, the method assumes a normal Gaussian distribution to the underlying data (which is often not the case in HCS) and is sensitive to outliers or otherwise non-symmetric distributions. For an approach which is more robust against outliers, the robust z-score uses the median for the mean and the median absolute deviation (MAD) for the standard deviation.
- **Robust linear scaling**: In this method, distributions are normalized by mapping the 1st percentile to 0 and the 99th percentile to 1[21]. This approach does not make assumptions about similarity of the distribution shape. However, this approach does not guarantee that the distributions of the controls will be identical between plates.

4.2 Intra-plate normalization

On a given plate, systematic errors generate repeatable local artifacts and smooth global drifts. These artifacts often becomes more noticeable upon visualization of the measurements and it is helpful to display well values graphically in their spatial layout, such as through the use of positional boxplots, heat maps of assay measurements on a plate layout and trellis map overviews of heat maps across multiple plates[23] (Figure 5).

In general, it is highly recommended to prospectively avoid such artifacts through sample preparation optimization. For example, plate edge effects can substantially mitigated by simply allowing newly cultured plates to incubate at room temperature for a period of time[14]. Another approach is to simply avoid the edge effect by leaving the edge wells filled with liquid but unused for samples; special plates exist for this purpose (e.g., Aurora plates from Nexus Biosystems). However, in experiments that require a large number of samples to be processed, leaving the edge wells empty may not be practical in terms of cost.

4.2.1 Correction of systematic spatial effects

The following analytical approaches may be used if there is an observed positional effect:

1. **Global parametric fitting**: This approach fits a smooth function to the data based on the physical plate layout. The corresponding per-cell measurements at each position are then divided by the smoothed function. Care must be taken in selecting the function parameters. If the function is too smooth, it is then unable to accurately model the spatial variation due to systemic error; if the function is too rigid, it will

over-fit the measurements and will not generalize to new data. Using splines to create the parametric surface is common.

2 **Local filtering**: Similar to the global parametric fitting approach, this method takes the physical plate layout into account. Assuming that the aberration is highly spatially localized, the measurements for each well are "denoised" using measures from adjacent wells, often the median calculated from a square neighborhood centered on the well to be normalized.

3 **B-Score**: This method locally corrects for systematic positional effects by iterative application of the Tukey median polish algorithm[19][24]. This approach is robust to outliers. However, it assumes that most samples in a plate have no biological effect (essentially using the entire plate as a negative control) and can produce artifacts if this assumption is violated.

4 **Model-based**: In either of the above cases, consideration must be given to whether all the samples on a plate can be used or just a subset. In a primary screen where the majority of the reagents can be assumed to have negligible or minor effect, the full set of samples can be used. In a confirmation screen, the spatial variation may be caused by samples that exhibit a moderate to large effect; hence, the representative samples should be drawn from the control wells. In this case, correction may be achieved using a diffusion model based on the control wells even though the location of the controls is often spatially constrained[25].

4.2.2 Illumination correction

The quantification of cellular fluorescence intensities and accurate segmentation of images is often hampered by variations in the illumination produced by the microscope optics and light sources. It is not uncommon for illumination to vary 1.5- to 2-fold across a single image when using standard microscope hardware.

Image acquisition software may mitigate these artifacts through the use of a reference image of a uniformly fluorescent sample (e.g. free fluorescent dye), which is then divided or subtracted from each collected image. This approach is described in "Image Optimization and Background Correction" section in the introductory HCS chapter and has the advantage of convenience and utility in cases where the illumination might change over time, e.g., light source aging. Disadvantages of this approach include:

- The underlying assumption is that the screener properly creates the appropriate referencing images and that conditions do not change between the acquisition of the reference images and the collection of the experimental images.
- Some software only allows one white reference image to be used to correct all wavelengths, which ignores differences in their respective optical paths and spectral characteristics.
- The fact that a uniformly fluorescent control is in a different chemical environment than a real sample is not taken into account.
- Typical methods do not account for different exposure times between the standard image and each collected image, although more sophisticated software allows for using a linear fit based on a series of images using a variety of exposure times for each wavelength.

A retrospective (i.e., post image acquisition) approach is an alternative[26]. It bases the correction on all (or a sufficiently large subset of) the images from a plate for a given wavelength. This approach assumes that the actual cellular intensity distribution is distorted by a multiplicative non-uniform illumination function (additional sources of bias may also be considered if needed). This approach is applied to each wavelength condition since the

spectral characteristics of the filters differ in addition to non-uniformities introduced by the excitation lamp. Furthermore, it should be applied to each plate for a given image acquisition run unless it can be shown that observed patterns are consistent across plates. The methodology is as follows:

1. For all images of a given wavelength, smooth the image by applying a median filter. Since we do not want the small-scale cellular features to obscure the underlying large-scale illumination heterogeneity, the size of the filter should large enough that any cells in the image are heavily blurred.
2. Estimate the illumination function by calculating the per-pixel average of all the smoothed images.
3. Rescale the illumination function so that the range is [1, ∞] by dividing by the minimum pixel value, or more robustly, by the 2ndpercentile pixel value (to avoid division-by-zero problems)
4. Obtain the corrected image by dividing the original image by the illumination correction function (Figure 6).

5 Measurement of image features

The quantitative extraction and measurement of features is performed by biological image analysis tools (e.g., CellProfiler[27], Fiji[28] and commercial software sold with HCS instruments). In addition to features that are straightforwardly related to the intended biological question, the extraction of additional features lends itself to serendipitous discoveries if mined correctly. A couple of examples illustrate this point:

- A phenotypic screen of 15 diverse morphologies in *Drosophila* cells revealed that cells with actin blebs and actin located in the periphery also tended to contain a 4N DNA content, a cell cycle relationship that would most likely not have been uncovered outside of an HCS context[29].
- Another study used a diffuse GFP reporter to look for clathrin-coated pit (plasma membrane) and intracellular vesicle formation. The researchers also took the opportunity measure GFP signal representing translocation to the nucleus. Unexpectedly, such an effect was uncovered which was compound-specific effect and not previously described in literature[30].

Measured image features fall into the following types[31] (some of which are described under "Feature Extraction" in the introductory HCS chapter):

- **Counts:** The number of objects or sub-cellular objects per compartment (e.g., foci per cell in a DNA damage experiment). The number of objects per image may also be useful as a quality control statistic.
- **Size and shape**: Size is a descriptor of the pixel area occupied by a specific labelled compartment and includes such measures as area, perimeter and major axis length. Shape measures describe specific spatial features. Some examples include the aspect ratio (ratio of the height vs width of the smallest enclosing rectangle) as a measure of elongation, or compactness (ratio of the squared perimeter and the area). Zernike features (coefficients of a Zernike polynomial fit to a binary image of an object) are also useful as descriptors of shape.
- **Intensity**: The amount of a particular marker at a pixel position, assuming that the total intensity is proportional to the amount of substance labeled. The idea is that the presence or absence of the marker reflects a specific cellular state. For example, the total intensity of a DNA label in the nucleus is related to DNA content, which is useful for cell-cycle phase identification. Intensity measurements include the minimum, maximum, various aggregate statistics (sum, mean, median) as well as

correlation coefficients of intensity between channels with different markers (useful for co-localization). If the target marker changes position, e.g. translocation from the nucleus from the cytoplasm, then the correlation coefficient of the stain between the two sub-compartments can provide a larger signal dynamic range than only measuring intensity in either the sub-compartment independently.

- **Texture**: A description of the spatial regularity or smoothness/coarseness of an object, which is useful for characterizing the finer patterns of localization. Textural features can be statistical (statistical properties of the grey levels of the points comprising a surface), structural (arrangements of regular patterns), or spectral (periodicity in the frequency domain).
- **Location**: The position of an object with respect to some other feature. Typically, the (x,y) location of an object within the image is not itself biologically relevant. However, relative positional features (e.g., absolute distance of foci from the border of an enclosing organelle) may be indicative of some physiological change.
- **Clustering**: A description of the spatial relationship of an object with respect to other objects. Examples of measures include the percentage of the perimeter touching a neighboring object and the number of neighbors per object.

It should be noted that while some of the suggested features are often difficult to use as direct readouts and are not biologically intuitive to interpret (e.g., texture and some shape features), they are often beneficial for machine learning approaches (see the section "Machine learning for HCS" below for more details). While spreadsheets are commonly used for storing cellular measurements, in order to contain the vast amount of per-cell information, across millions of cells, a database is a more feasible option for data storage and interrogation.

6 Machine learning for HCS

Generically, *machine learning* is defined as the use of algorithms capable of building a model from existing data as input and generalizing the model to make predictions about unknown data as output[32]. The measurement of a large number of features lends itself to the use of machine-learning approaches for automated scoring of samples, especially in cases where visual inspection or hand-annotation of images is time- and cost-prohibitive. In the context of HCS, machine-learning algorithms make predictions about images (or regions of images) based on prior training. We describe below two domains in HCS where machine-learning has proven useful: identifying regions of interest in images and scoring phenotypes (a brief discussion of machine-learning for quality control application is above).

6.1 Machine learning for image segmentation

Typically, the first step to an HCS workflow is the identification of the image foreground from the background, i.e, finding which pixels belong to each object of interest. For fluorescent images, often one of a number of thresholding algorithms is suitable for this purpose[33].

However, in cases where the pixel intensity of the foreground is not markedly different from that of the background (e.g, brightfield images), machine-learning approaches can be useful in classifying pixels as foreground and background based on other features, such as local intensity variation or texture. Since it not trivial to choose *a priori* features that identify the foreground, a common approach is to extract a large number of image features, hand-select example foreground and background regions and then use machine learning to find combinations of features that identify the foreground class (or classes) of pixels from the background.

One open-source pixel classification tool is Ilastik[34]. It classifies each pixel in an image by calculating sets of features that are linear combinations of the intensities of neighborhood pixels in order to identify textures and edges in the neighborhood of the pixel[35]. The software then calculates a membership probability for each class based on these features using supervised machine learning based on hand-annotated pixels.

6.2 Machine learning for scoring phenotypes

Machine-learning can in theory be used to identify samples of interest based on features calculated from entire images. However, the actual application of this approach is rare in HCS, so we focus more on machine-learning based on per-cell features.

6.2.1 Feature extraction and normalization

Using one or two features for scoring phenotypes is a common approach, especially when the biological relevance of the features of interest are well-defined[17]. However, hand-selecting the features necessary to distinguish phenotypes of interest versus negative controls is often intractable, especially if the phenotype is subtle, or simple linear combinations are insufficient. A machine learning approach lends itself well to this task, provided a sufficient variety of features are provided as "raw material." For more details on the types of features that can be extracted see the section "Measurement of image features" above. For complex phenotypes, the features that will contribute to the discriminating power of the classifiers will probably not be known in advance. Thus, in general, it is advisable to extract as many features as is practical and to use a machine-learning algorithm capable of choosing among them.

In many HCS experiments, the phenotype is dramatic enough that plate-to-plate variation and batch effects are the only confounding effects of concern and they can be removed adequately using the techniques mentioned in the prior section.

However, for more subtle phenotypes, the screener will need to model and remove the confounding effects more precisely in order to avoid obscuring the distinctions between phenotypes of interest.

For further details related to normalization, see the "Normalization" section above.

6.2.2 Supervised machine-learning for identification of particular phenotype sub-populations

For rare phenotypes that are nonetheless recognizable by eye, a researcher can generate a classifier to recognize cells with the phenotype of interest. Software packages that perform this task are Definiens Cellenger[36] and CellProfiler Analyst[29]. Here we describe the workflow for CellProfiler Analyst, which is open-source (available at http://www.cellprofiler.org).

6.2.2.1 Creating a classifier

In CellProfiler Analyst, an interactive training session with iterative feedback is used to create a classifier for supervised machine learning as follows:

1. The software presents cells to the researcher for sorting, either selected randomly from the assay or taken from a specific plate with positive or negative controls. The screener manually sorts these into phenotypic "bins" to create a *training set.* Preferably, the screener will sort clear examples of the phenotype(s) in question; cells with an uncertain phenotype can be ignored, while keeping in mind that all cells will eventually be scored by the computer. Here, we refer to cells showing a phenotype of interest as "positives" (Figure 7).

Additional bins can be added, but as few bins as necessary should be used for the relevant downstream analysis as adding too many bins can decrease the overall accuracy.

If uncertain about the classification of a particular object, it can be ignored or removed from the list of objects under consideration. However, keep in mind that the final scoring will ultimately assign **all** objects to a class.

A note of caution: Sampling of a phenotype from only the control samples can lead to "overfitting," a scenario in which the machine-learning algorithm preferentially learns features which are irrelevant to the phenotype itself (e.g., spatial plate effects), leading to a classifier which does not generalize well and has poor predictive performance. It is thus preferred to select individual cells from a variety of images in the experiment.

2 After enough initial examples are acquired for the training set (typically, a few dozen or so), the screener then requests the computer to generate a tentative classifier based on the sorted cells.The screener sets the number of rules for distinguishing the cells in each of the classification bins. Here, we define a *rule* as a feature and cutoff representing a decision about the cell. It is recommended to use a smaller number of rules (e.g., five) at the early stages of defining a training set in order to accumulate cells spanning the full range of the phenotype and avoid training for a too-narrow definition of the phenotype (Figure 8).

3 At this point, the goal is to refine the rules by adding more cells to the training set. The screener then requests cells that the classifier deems as belonging to a particular phenotype.

4 The screener refines the training set, correcting errors by moving misclassified cells to the correct bin and re-training the classifier by generating a new set of rules (Figure 9).

5 By repeating the two steps above, the classifier becomes more accurate. If needed, the number of rules should be increased to improve accuracy (see below). At this point, the screener should save the training set for future refinement, to re-generate scores and for experimental records. It is advisable to do so before proceeding to scoring the entire experiment since scoring may take a long time for large screens.

6 When the accuracy of the classifier is sufficient, the screener can then scores all cells in the experiment so that the number of positive cells in each sample can be calculated (Figure 10).

6.2.2.2 Obtaining rules during the training phase and assessing classier accuracy

When the first pass at sorting sample cells is finished, the maximum number of rules allowed needs to be specified prior to generating the initial set of rules.

During the initial training step, it is best to use a small number of rules (typically 5 to 10) in order to avoid defining the phenotype too narrowly. Doing so will help insure identification of the minimal set of features covering the wide range of object characteristics represented in the training set.

As training proceeds, if the number of misclassifications does not improve, the number of rules may be increased to allow the machine-learning algorithm to capture more subtle distinctions between phenotypes. However, using more rules does not always result in

greater accuracy. In particular, increasing the number of rules above 100 is unlikely to improve classification accuracy and is computationally expensive to calculate.

Based on prior experience with 14 phenotypes in human cells, an upper limit of 50 rules is recommended for complex object classes (that is, to the human eye, one that involves the assessment of many features of the objects simultaneously)[29].

The most accurate way to gauge the performance of a classifier is to fetch a large number of objects of a given phenotype from the whole experiment. The fraction of the retrieved objects correctly matching the requested phenotype indicates the classifier's general performance. For example, if a screener fetches 100 positive objects but find upon inspection that 5 of the retrieved objects are not positives, then the classifier is estimated to have a positive predictive value of 95% on individual cells. Note that the classifiers' ability to detect positives and negatives must be interpreted in the context of the actual prevalence of individual phenotypes, which may be difficult to assess a priori. For studies or screens where the data is gathered over time, re-training the classifiers on the larger data set can increase their robustness.

Cross-validation is a standard method for estimating classifier accuracy, with important caveats discussed below. One version of this approach is to use a sub-sample of the training set for training a classifier and then use the remainder of the training set for testing. The optimal number of rules may be assessed by plotting the cross-validation accuracy for the training set as an increasing number of rules are used, where values closer to 1 indicate better performance. Two features of the plot are useful for guiding further classification:

1. If the accuracy increases (that is, slopes upward) at larger numbers of rules, adding more rules is likely to help improve the classifier (if the line slopes downward, this may indicate more training examples are needed).
2. If the accuracy is displayed for two sub-sampling percentages (say, 50% and 95% of the examples are used for training), and the two curves are essentially the same, adding more cells to the training set is unlikely to improve performance.

A note of caution: The accuracy in these plots should not be interpreted as the actual accuracy for the overall experiment. These plots tend to be pessimistic, as the training set usually includes a disproportionate number of difficult-to-classify examples (Figure 11).

The relationship between accuracy on individual cells versus accuracy for scoring wells for follow-up is complicated, because false positives and false negatives are often not evenly distributed across wells in an experiment. In practice, improving accuracy on individual cells leads to better accuracy on wells, and in general, the actual goal is per-well accuracy more than per-cell accuracy.

7 Whole-organism HCS

For those experiments in which the molecular mechanisms in question cannot yet be reduced to biochemical or cell-based assays, screeners can search for chemical or genetic regulators of biological processes in whole model organisms rather than isolated cells or proteins. The advantages of performing HCS in an intact, physiological system include the increase in likelihood that the findings from such experiments accurately translate into the context of the human body (e.g., in terms of toxicity and bioavailability), simplification of the path to clinical trials, and reduction of the failure of potential therapeutics at later stages of testing.

While a number of small animals are amenable to whole-organism HCS (e.g., zebrafish embryos, *Drosophila* fruit fly larvae, etc), this section of the chapter will focus on novel HCS techniques developed for the *Caenorhabditis elegans* roundworm.

7.1 *C. elegans* HCS

The nematode *C. elegans* is a increasingly popular choice for enabling HCS in whole organisms due to the following advantages:

- Manually-analyzed RNAi and chemical screens are well-proven in this organism, with dozens completed[37].
- Many existing assays can be adapted to HCS; instrumentation exists to handle and culture *C. elegans* in HTS-compatible multi-well plates.
- Its organ systems have high physiologic similarity and genetic conservation with humans.
- *C. elegans* is particularly suited to assays involving visual phenotypes: physiologic abnormalities and fluorescent markers are easily observed because the worm is mostly transparent.
- The worms follow a stereotypic development pattern that yields identically-appearing adults, such that deviations from wild-type are more readily apparent.

Microscopy imaging and flow cytometry are the primary HCS methods for *C. elegans*, as plate readers do not offer per-worm or morphological readouts and often cannot measure bulk fluorescence from worm samples due to their spatial heterogeneity within the well.

- **Flow cytometers**: Systems such as the COPAS Biosort (Union Biometrica) can be used for the automatic sorting of various "large" objects, including *C. elegans*, using the object size and intensity of fluorescent markers. Such a system is capable of differentiating some phenotypes using fluorescent intensity changes as the readout (e.g., isolating of mutants with reduced RFP-to-GFP intensity ratios as compared to wild-type worms[38] or signature extraction based on GFP intensity profiles created along the length of the worm[39]). One limitation of this approach is low spatial resolution. Another disadvantage is that retrieval of worms from the multi-well plates typically used in screening becomes a rate-limiting step, reducing the throughput.
- **Automated microscopy**: Defining worm phenotypes in HCS has also been enabled through the use of image data. Standard 6- or 12-well assays can be miniaturized to 96- or 384 well plates by dispensing a precise number of worms within a specified size/age range into the desired number of multiwell plates (using a COPAS sorter, for example). The worms are typically transferred from agar to liquid media to minimize imaging artifacts. A paralytic drug may be added to slow worm movement, minimizing misalignment between subsequently imaged channels. Alternately, microfluidics may also be used to stabilize worm position[40].

Below is an HCS image analysis workflow tailored for worms dispensed into multi-well plates[41], using brightfield images of each well, along with the corresponding images from additional fluorescence wavelengths:

1. **Well identification:** In order to restrict worm detection to the region of interest, delineate the well boundary within the brightfield image; the well interior is typically brighter than the well exterior, lending itself to simple thresholding. In order to avoid artifacts at the well edge, use morphological erosion to contract the well border by a few pixels.
2. **Illumination correction and masking of pixel intensities:** Often the brightfield image will exhibit illumination heterogeneities which must be corrected prior to worm detection (see the "Illumination_correction" section above). Calculate an illumination correction function (ICF) from the brightfield image and then correct the image by dividing by the ICF. It is recommended to mask the image using the

eroded well image and use the result for creating the ICF, otherwise the ICF will be distorted by the sharp features at the well edge. At this point, the illumination-corrected image is then masked with the eroded well image in order to restrict worm identification to the well area.

3 **Worm foreground identification:** Identify the worms as the image foreground using image thresholding on the brightfield image. Since the brightfield images are usually high-contrast, an automatic thresholding method such as Otsu is typically effective here. It is helpful to impose size criteria in order to remove objects that are likely to be spurious, e.g, debris, embryos, and other artifacts.

4 **Make population-averaged measurements if desired:** At this point, quantification of the additional fluorescent markers within the worm regions can be performed. For example, if a viability stain (e.g., SYTOX Orange) was used as part of a live/dead assay, image segmentation (i.e. partitioning the foreground pixels into individual worms) is not necessary, and the workflow would continue by identifying the SYTOX-positive pixels from the fluorescent image using automatic thresholding, measuring the total pixel area occupied by the worm foreground and the SYTOX foreground, and calculating the ratio of the SYTOX foreground total area and the worm foreground area to yield the final per-well readout of worm death.

5 **Make per-worm measurements if desired:** For some assays, it is preferable to identify individual animals rather than a whole-well readout (e.g., pathogen screens). While non-touching worms can usually be delineated in brightfield images based on the differences in intensities between foreground and background, image intensity alone is not sufficient for touching and overlapping worms. For these assays, algorithms are required that separate touching and overlapping worms. Moreover, edges and intensity variations within the worms often mislead conventional segmentation algorithms. Here, we describe a recent algorithm that employs a probabilistic shape model using intrinsic geometrical properties of the worms (such as length, width profile, and limited variability in posture) as part of an automated segmentation method (distributed as a toolbox in CellProfiler)[42] (Figure 12)

   a **Identify worms as described in the above workflow:** in this case, however, only the brightfield images are acquired from each well

   b **Construct a worm "model"**: Once the worm foreground is obtained, construct a model of the variations in worm morphology by creating a training set of non-touching representative worms. This is done by saving binary images of a number of non-touching worms and using the "Training" mode of the UntangleWorms module in the CellProfiler worm toolbox.

   c **Apply the worm model to worm clusters:** Once the model is created, apply the model on the images using the "Untangle" mode of the UntangleWorms module. The result of this operation will be identify the individual worms from the worm clusters as well as exclude artifacts such as debris, embryos, etc.

   d **Quantify fluorescent markers:** Measure the various features available for each worm, such as morphology, intensity, texture, etc. The delineated worms can also be mapped to a common atlas (using StraightenWorms) so that spatial distribution of staining patterns may be quantified (Figure 12).

8 Acknowledgements

This material is based upon work supported by the National Institutes of Health (R01 GM089652 to AEC, U54 HG005032 to Stuart Schreiber, and RL1 HG004671 to Todd Golub, administratively linked to RL1 CA133834, RL1 GM084437, and UL1 RR024924), the National Science Foundation (DB-1119830 to MAB), and Eli Lilly & Company.

9 References

1. Kozak K, Agrawal A, Machuy N, Csucs G. Data mining techniques in high content screening: A survey. (2009). J Comput Sci Syst Biol 2009;2:219–239.
2. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. Nat Biotechnol 2006;24(2):167–75. [PubMed: 16465162]
3. Genovesio A, Kwon YJ, Windisch MP, Kim NY, Choi SY, Kim HC, Jung S, Mammano F, Perrin V, Boese AS, Casartelli N, Schwartz O, Nehrbass U, Emans N. Automated genome-wide visual profiling of cellular proteins involved in HIV infection. J Biomol Screen. 2011;16(9):945–58. [PubMed: 21841144]
4. Zhang JH, Chung TDY, Oldenburg KR. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J Biomol Screen 1999;4(2):67–73. [PubMed: 10838414]
5. Iversen PW, Eastwood BJ, Sittampalam GS, Cox KL. A comparison of assay performance measures in screening assays: signal window, Z'factor, and assay variability ratio. J Biomol Screen 2006;11(3):247–52. [PubMed: 16490779]
6. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, Santoyo-Lopez J, Dunican DJ, Long A, Kelleher D, Smith Q, Beijersbergen RL, Ghazal P, Shamu CE. Statistical methods for analysis of high-throughput RNA interference screens. Nat Methods 2009;6(8):569–75. [PubMed: 19644458]
7. Sui Y, Wu Z. Alternative statistical parameter for high-throughput screening assay quality assessment. J Biomol Screen 2007;12(2):229–34. [PubMed: 17218666]
8. Ravkin I, Temov V, Nelson AD, Zarowitz MA, Hoopes M, Verhovsky Y, Ascue G, Goldbard S, Beske O, Bhagwat B, Marciniak H. Multiplexed high-throughput image cytometry using encoded carriers. Proc SPIE 2004;5322:52–63.
9. Zhang XHD. A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. Genomics 2007;89(4):552–61. [PubMed: 17276655]
10. Zhang XHD (2011). Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-scale RNAi Research. Cambridge University Press, 2011 ISBN 978-0-521-73444-8.
11. "Strictly standardized mean difference." Wikipedia, The Free Encyclopedia. Wikimedia Foundation,http://en.wikipedia.org/wiki/Strictly_standardized_mean_difference
12. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27(8):861–874.
13. "Receiver operating characteristic", Wikipedia, The Free Encyclopedia. Wikimedia Foundation,http://en.wikipedia.org/wiki/Receiver_operating_characteristic
14. Lundholt BK, Scudder KM, Pagliaro L. A simple technique for reducing edge effect in cell-based assays. J Biomol Screen. 2003;8(5):566–70. [PubMed: 14567784]
15. Bray MA, Fraser AN, Hasaka TP, Carpenter AE. Workflow and metrics for image quality control in large-scale high-content screens. J Biomol Screen. 2012;17(2):266–74. [PubMed: 21956170]
16. Rajaram S, Pavie B, Wu LF, Altschuler SJ. PhenoRipper: software for rapidly profiling microscopy images. Nat Methods. 2012;9(7):635–637. [PubMed: 22743764]
17. Logan DL, Carpenter AE. Screening cellular feature measurements for image-based assay development. J Biomol Screen. 2010;15(7):840–846. [PubMed: 20516293]
18. Dragiev P, Nadon R, Makarenkov V. Systematic error detection in experimental high-throughput screening. BMC Bioinformatics 2011;12:25. [PubMed: 21247425]
19. Brideau C, Gunter B, Pikounis B, Liaw A. Improved statistical methods for hit selection in high-throughput screening. J Biomol Screen 2003;8(6):634–47. [PubMed: 14711389]

20. Snijder B, Pelkmans L. Origins of regulated cell-to-cell variability. Nat Rev Mol Cell Biol 2011;12(2):119–25. [PubMed: 21224886]
21. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. Science 2004;306(5699):1194–8. [PubMed: 15539606]
22. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80. [PubMed: 15461798]
23. Gunter B, Brideau C, Pikounis B, Liaw A. Statistical and graphical methods for quality control determination of high-throughput screening data. J Biomol Screen. 2003;8(6):624–33. [PubMed: 14711388]
24. Makarenkov V, Zentilli P, Kevorkov D, Gagarin A, Malo N, Nadon R. An efficient method for the detection and elimination of systematic error in high-throughput screening. Bioinformatics. 2007;23(13):1648–57. [PubMed: 17463024]
25. Carralot JP, Ogier A, Boese A, Genovesio A, Brodin P, Sommer P, Dorval T. A novel specific edge effect correction method for RNA interference screenings. Bioinformatics 2012;28(2):261–8. [PubMed: 22121160]
26. Jones TR, Carpenter AE, Sabatini DM, Golland P (2006) Methods for high-content, high-throughput image-based cell screening. Proceedings of the Workshop on Microscopic Image Analysis with Applications in Biology (MIAAB). Metaxas DN, Whitaker RT, Rittcher J, Sebastian T (Eds). Copenhagen, Denmark, October 5, pp 65-72.
27. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006;7(10):R100. [PubMed: 17076895]
28. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9(7):676–82. [PubMed: 22743772]
29. Jones TR, Carpenter AE, Lamprecht MR, Moffat J, Silver SJ, Grenier JK, Castoreno AB, Eggert US, Root DE, Golland P, Sabatini DM. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. Proceedings of the National Academy of Sciences USA 2009;106(6):1826–31.
30. Garippa RJ, Hoffman AF, Gradl G, Kirsch A. High Throughput Confocal Microscopy for Beta Arrestin Green Fluorescent Protein Translocation: G Protein-Coupled Receptor Assays Using the Evotec (Perkin-Elmer) Opera. Methods in Enzymology: Measuring Biological Responses with Automated Microscopy, Vol. 414. Ed. By James Inglese, Elsevier Academic Press, San Diego, CA. 2006.PMID 17110189
31. Ljosa V, Carpenter AE. Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening. PLoS Comput Biol 2009;5(12):e1000603. [PubMed: 20041172]
32. Tarca AL, Carey VJ, Chen XW, Romero R, Drăghici S. Machine learning and its applications to biology. PLoS Comput Biol. 2007;3(6):e116. [PubMed: 17604446]
33. Sahoo PK, Soltani S, Wong AK, Chen YC. A survey of thresholding techniques. Comput Vision Graph Image Process 1988;41:233–260.
34. ilastik: Interactive Learning and Segmentation Toolkit.http://www.ilastik.org/
35. Hamprecht FA (2010) Ilastik: Interactive learning and segmentation tool kit. 2011 IEEE International Symposium on Biomedical Imaging. URL http://hci.iwr.uni-heidelberg.de/download3/ilastik.php.
36. Baatz M, Arini N, Schäpe A, Linssen B. Object-oriented image analysis for high content screening: detailed quantification of cells and sub cellular structures with the Cellenger software. Cytometry A 2006;69(7):652–8. [PubMed: 16680706]

37. O'Rourke EJ, Conery AL, Moy TI. Whole-animal high-throughput screens: the C. elegans model. Methods Mol Biol. 2009;486:57–75. [PubMed: 19347616]

38. Doitsidou M, Flames N, Lee AC, Boyanov A, Hobert O. Automated screening for mutants affecting dopaminergic-neuron specification in C. elegans. Nat Methods 2008;5(10):869–72. [PubMed: 18758453]

39. Dupuy D, Bertin N, Hidalgo CA, Venkatesan K, Tu D, Lee D, Rosenberg J, Svrzikapa N, Blanc A, Carnec A, Carvunis AR, Pulak R, Shingles J, Reece-Hoyes J, Hunt-Newbury R, Viveiros R, Mohler WA, Tasan M, Roth FP, Le Peuch C, Hope IA, Johnsen R, Moerman DG, Barabási AL, Baillie D, Vidal M. Genome-scale analysis of in vivo spatiotemporal promoter activity in Caenorhabditis elegans. Nat Biotechnol 2007;25(6):663–8. [PubMed: 17486083]

40. Chung K, Crane MM, Lu H. Automated on-chip rapid microscopy, phenotyping and sorting of C. elegans. Nat Methods. 2008;5(7):637–43. [PubMed: 18568029]

41. Moy TI, Conery AL, Larkins-Ford J, Wu G, Mazitschek R, Casadei G, Lewis K, Carpenter AE, Ausubel FM. High-throughput screen for novel antimicrobials using a whole animal infection model. ACS Chemical Biology 2009;4(7):527–33. [PubMed: 19572548]

42. Wählby C, Kamentsky L, Liu ZH, Riklin-Raviv T, Conery AL, O'Rourke EJ, Sokolnicki KL, Visvikis O, Ljosa V, Irazoqui JE, Golland P, Ruvkun G, Ausubel FM, Carpenter AE. An image analysis toolbox for high-throughput C. elegans assays. Nat Methods 2012;9(7):714–6. [PubMed: 22522656]

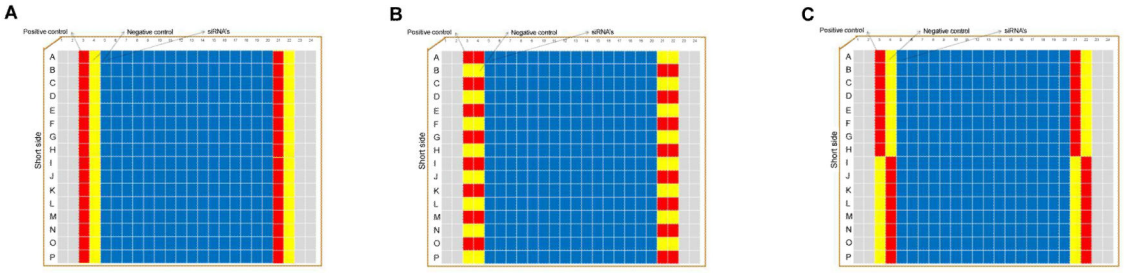

Figure 1: Location of sixteen positive controls (red) and sixteen negative controls (yellow) on a 384-well plate. In layout (A), both sets of controls are located on the plate edges in a regular pattern, and are susceptible to edge-based bias. In contrast, layouts (B) and (C) attempt to systematically decrease the edge bias by alternating the spatial position of the controls so that they appear in equal quantity on each of the rows and available columns. Adapted from [1], copyright OMICS Publishing Group..

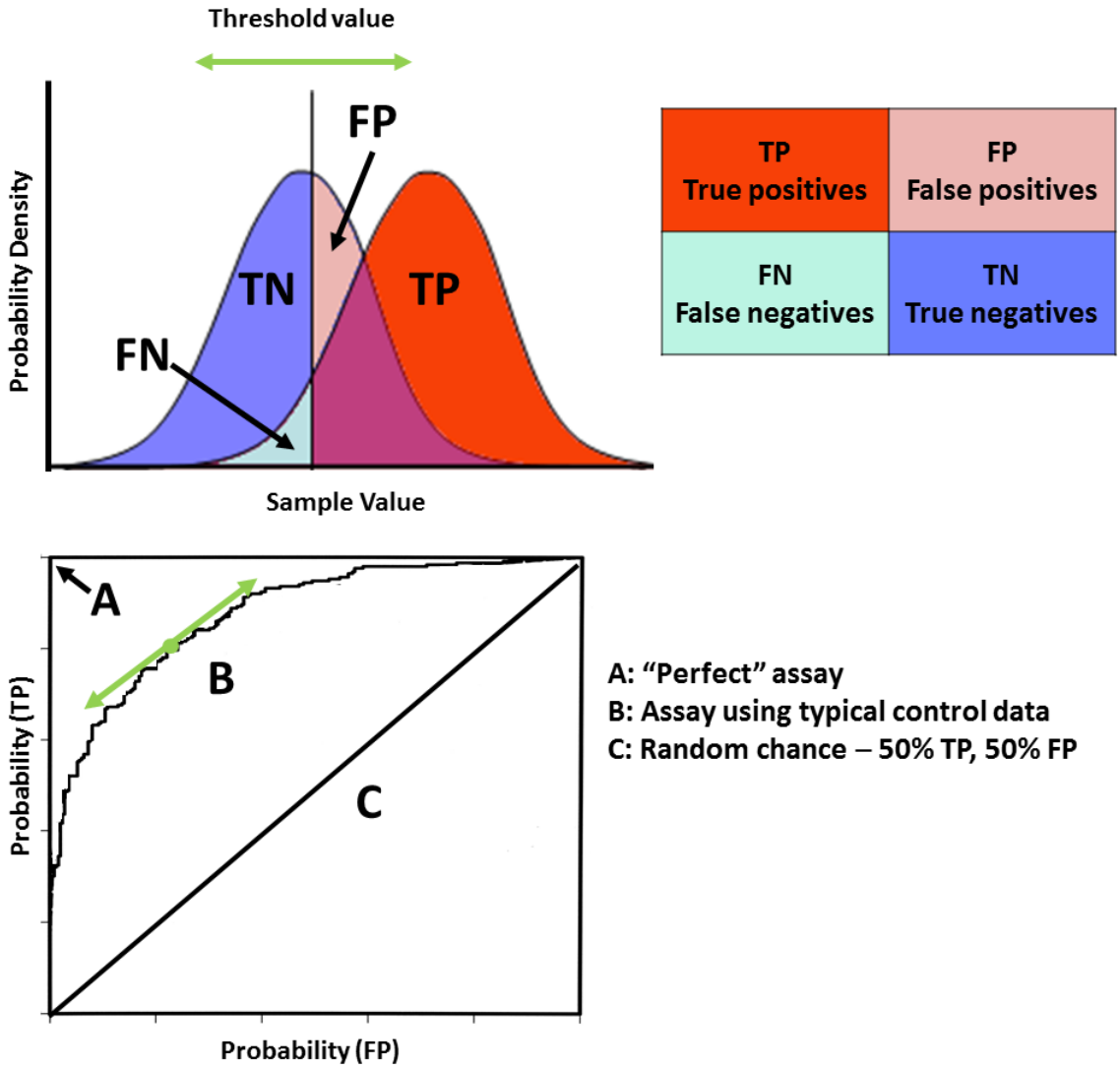

Figure 2: Details of a receiver characteristic curve. "Top:" Probability distributions of a hypothetical pair of control data. As a threshold value is varied, the proportions of actual and predicted positives and negatives drawn from the two distributions will also vary. "Bottom:" Three hypothetical ROC curves based on Table 3. Adapted from [10].

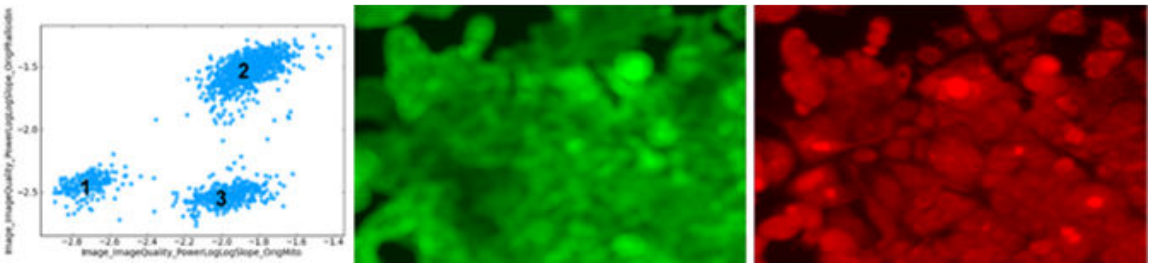

Figure 3: Illustration of PLLS performance for in-focus/out-of-focus MCF7 images. *Left panel*: Scatter plot of PLLS from the mitochondrial channel (x-axis) vs the phalloidin channel (y-axis) on a whole-image basis. If both channels were simultaneously blurred, we would expect that the measurements would cluster along a line. In this case, though, there are three clusters are

Assay Guidance Manual

Assay Guidance Manual

apparent: clusters 1 and 2 represent images in which both channels are blurred/in-focus, and cluster 3 where one channel is in focus but the other is not. An example image from cluster 3 is shown, both the blurry mitochondrial (center panel) and in-focus phalloidin channels (right panel).

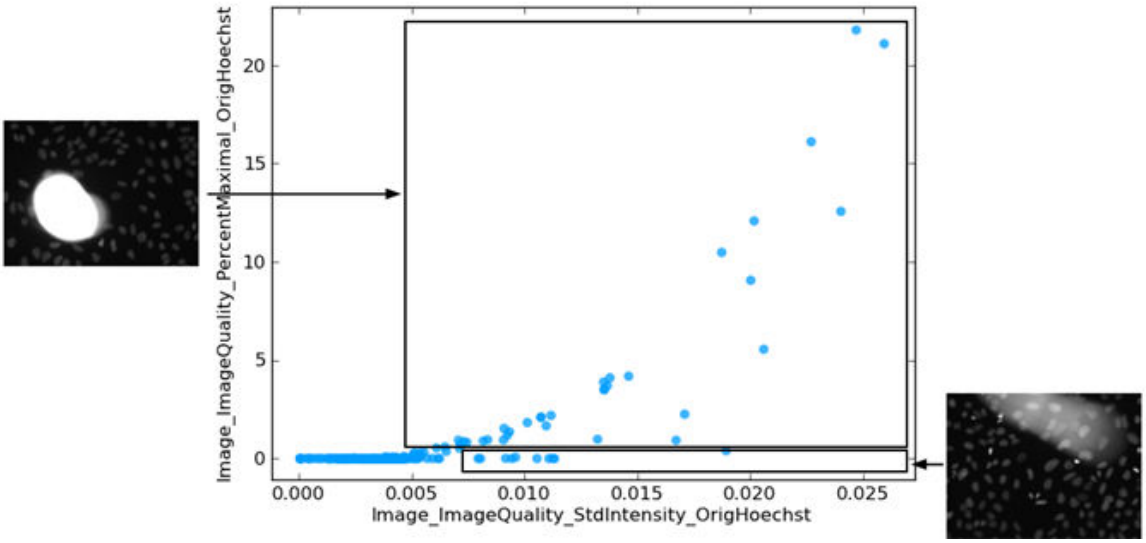

Figure 4: Illustration of performance of the percentage of pixels at maximum intensity and the standard deviation of the pixel intensities. The inset to the left shows a example image from the box at top where the percentage measure is high. The inset to right shows an example image from the box at bottom where the image standard deviation is high but the percentage measure is low.

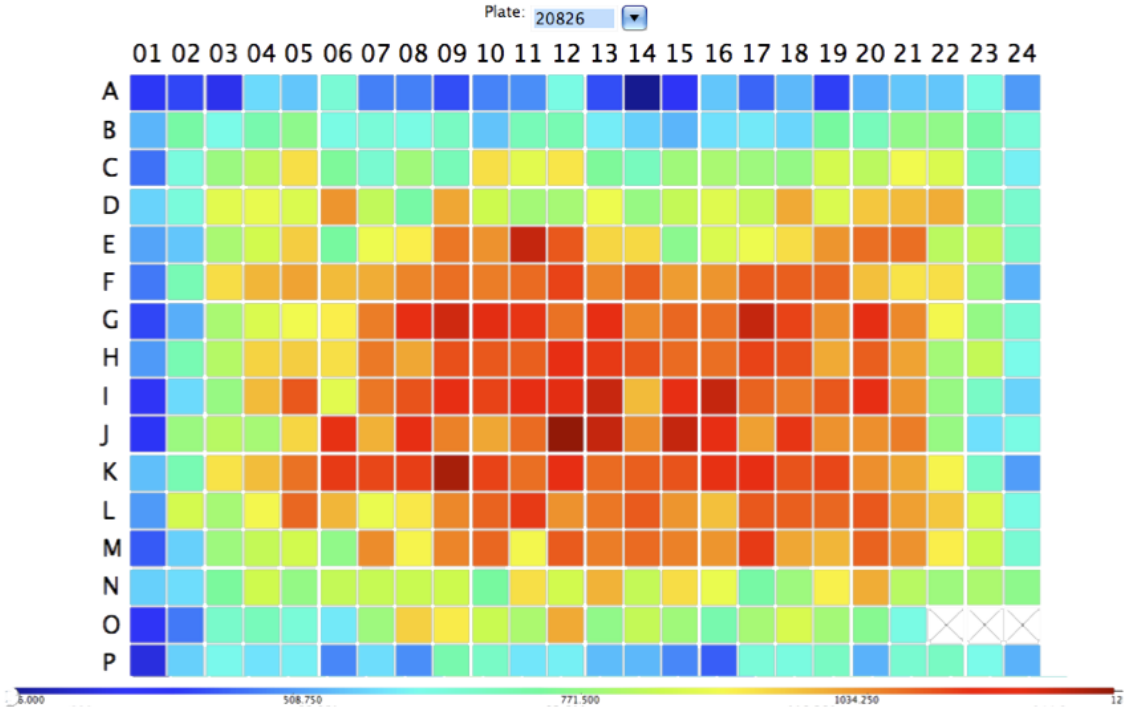

Figure 5: Heat map of values from a hypothetical 384-well plate containing edge effects.

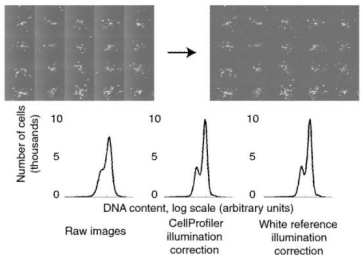

Figure 6: *Top left:* Example of uneven illumination from the left to the right within each field of view in a tiled grid of 5 x 4 images from a cell microarray. *Top right:* Correction of anomalies by CellProfiler. *Bottom:* Impact of anomalies and correction on Drosophila Kc167 DNA content data. Adapted from [27], copyright Carpenter et al.

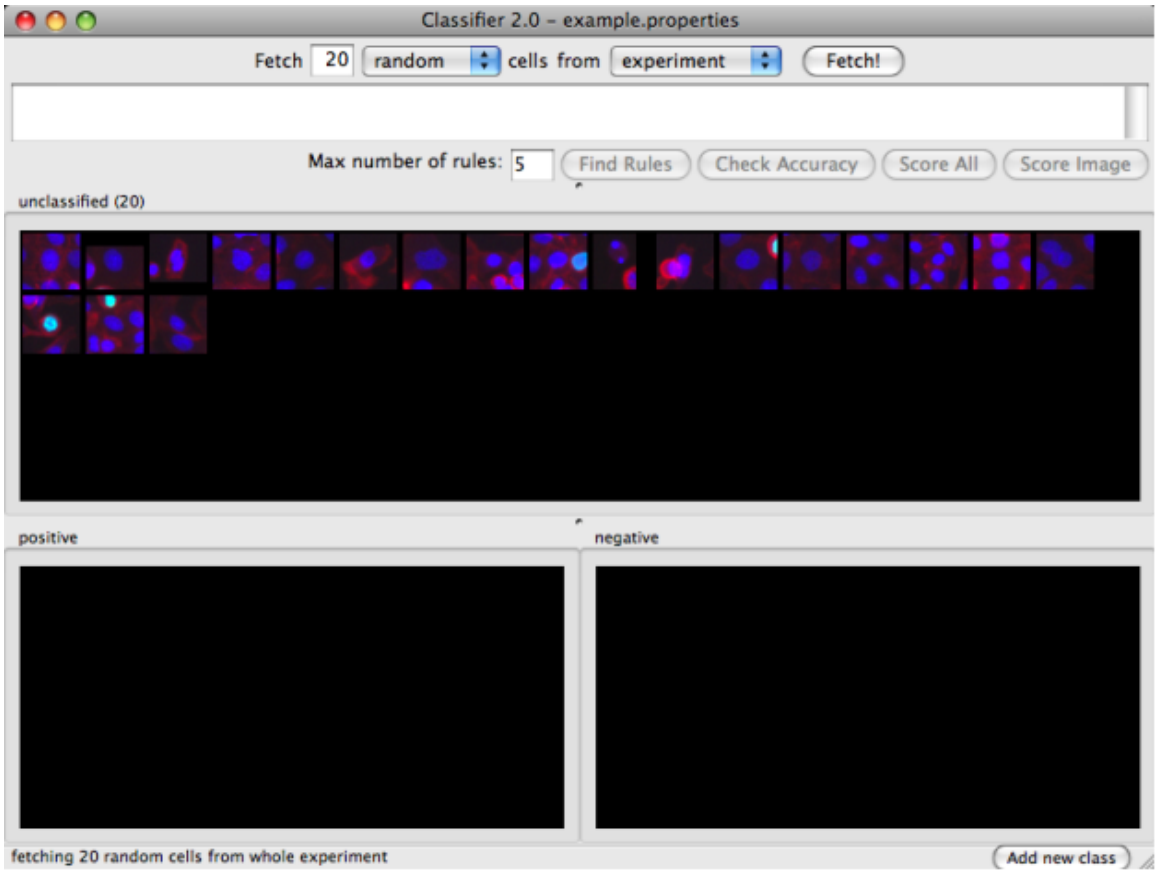

Figure 7: Twenty unclassified cells are presented to the screener for initial sorting using the CellProfiler Analyst phenotype classification tool.

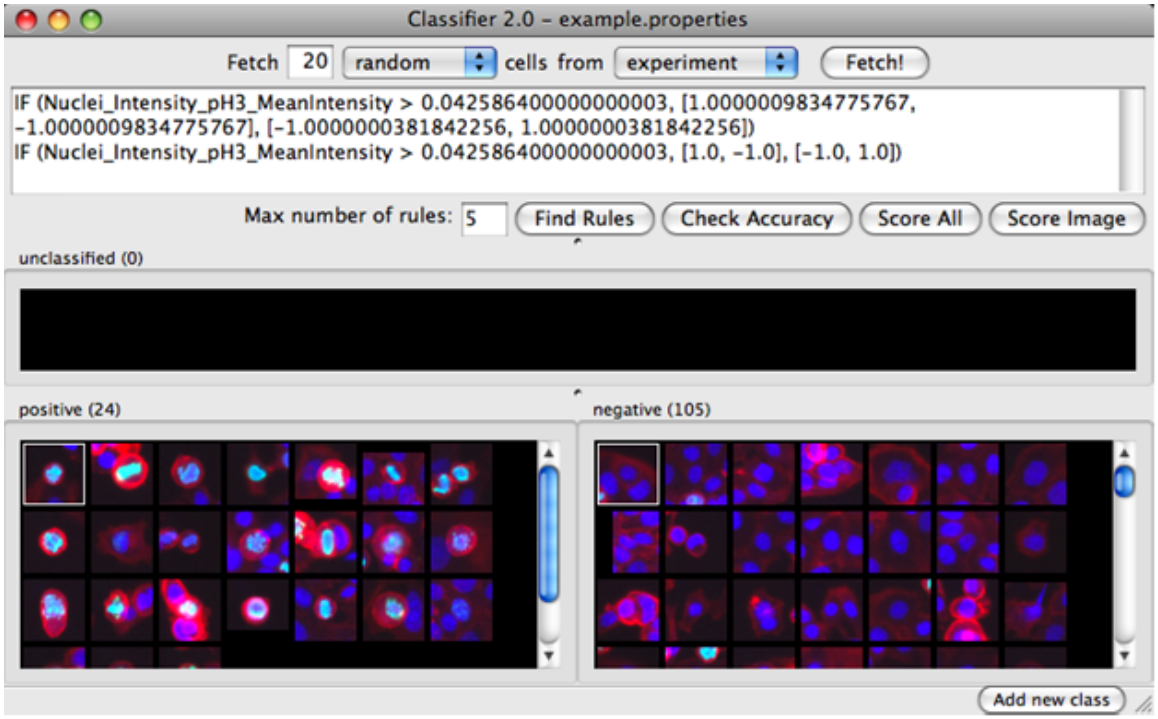

Figure 8: The set of rules after initial sorting using the CellProfiler Analyst classifier tool. In this example, only 2 rules were found out of the specified maximum of 5, both pertaining to the mean pH3 intensity of the nuclei channel, indicating that this feature was sufficient to achieve perfect classification on the training set.

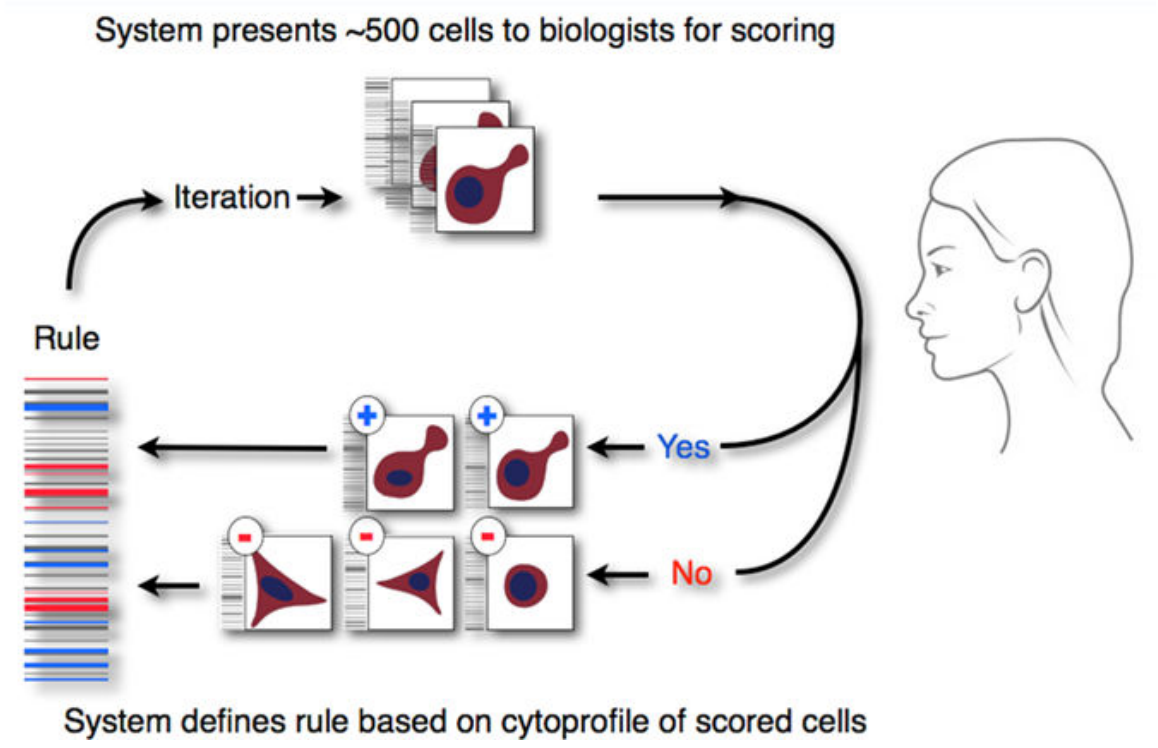

Figure 9: Illustration of the iterative machine learning workflow. Adapted from [29], copyright The National Academy of Sciences of the USA.

Assay Guidance Manual

Assay Guidance Manual

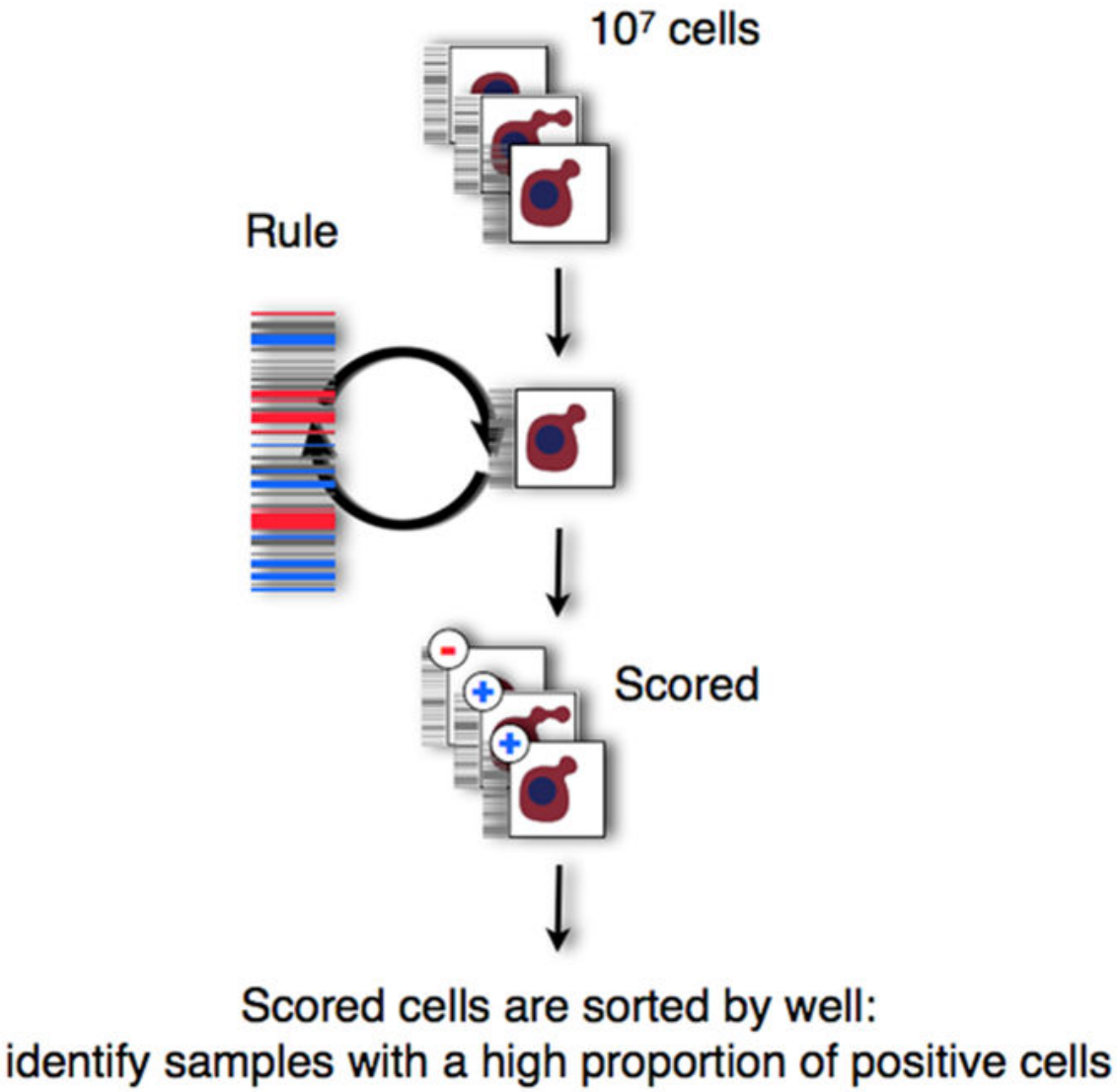

Figure 10: Illustration of the final cell scoring workflow. Adapted from [29], copyright The National Academy of Sciences of the USA.

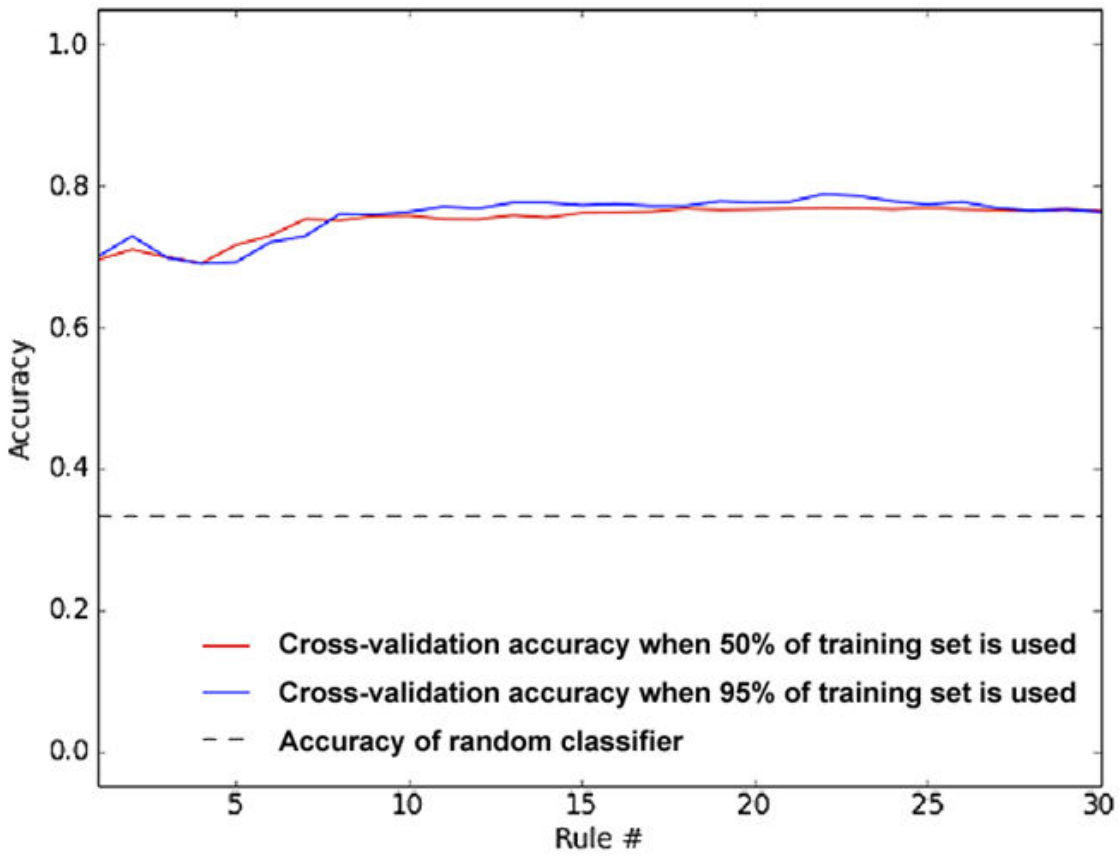

Figure 11: Plot displaying the cross-validation accuracy of a 3-class classifier with 30 rules. Note that the accuracy does not increase for more than 10 rules.

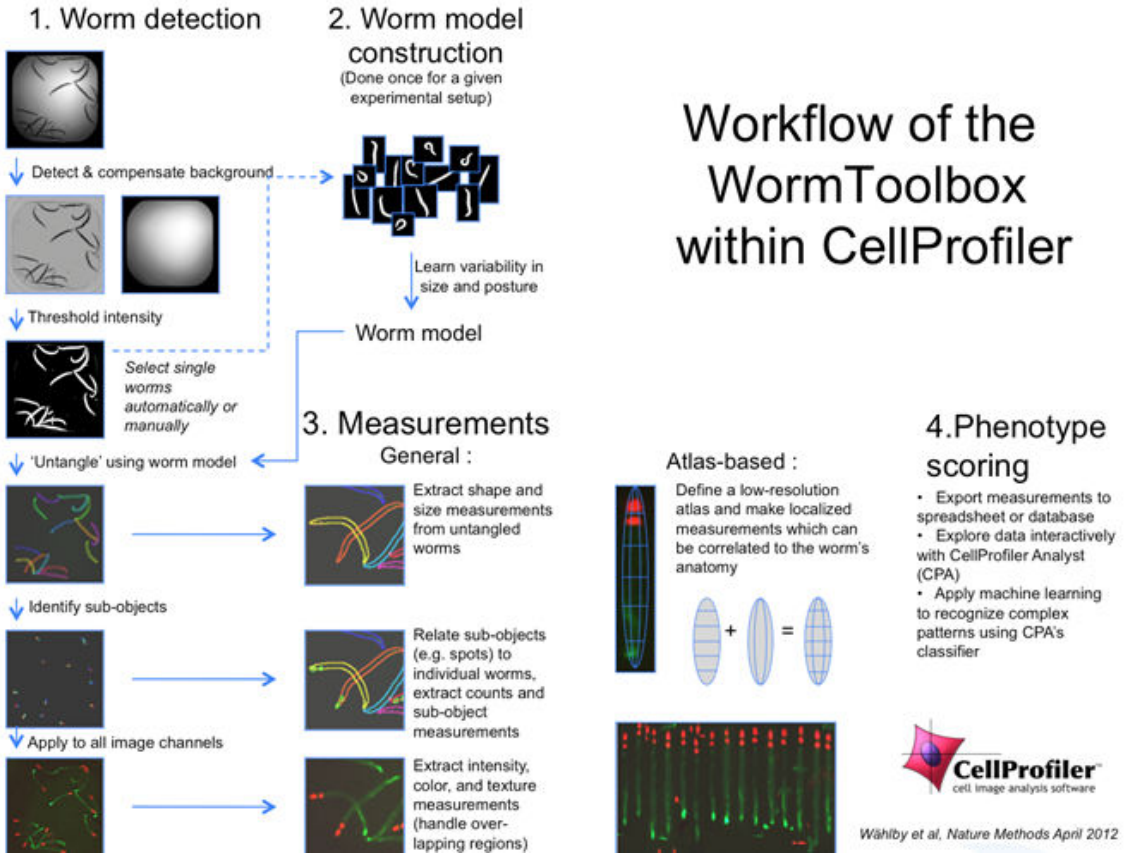

Figure 12: Workflow of the WormToolbox in CellProfiler. Adapted from [42], contributed by Carolina Wählby.

Assay Guidance Manual

Assay Guidance Manual

Table 1: Interpretation of Z’ values

Value	**Interpretation**
$Z' = 1$	"Perfect"
$Z' \geq 0.5$	Excellent: Good separation between the populations
$0.5 < Z' \leq 0$	Acceptable: Moderate separation of the distributions
$Z' = 0$	Nominal: Good only for a yes/no response
$Z' < 0$	Unacceptable

Assay Guidance Manual

Assay Guidance Manual

Table 2: Interpretation of SSMD values (positive control response > negative control response)

Quality Type	**(1) Moderate Control**	**(2) Strong Control**	**(3) Very Strong Control**	**(4) Extremely Strong Control**
Excellent	$\beta \geq 2$	$\beta \geq 3$	$\beta \geq 5$	$\beta \geq 7$
Good	$2 > \beta \geq 1$	$3 > \beta \geq 2$	$5 > \beta \geq 3$	$7 > \beta \geq 5$
Inferior	$1 > \beta \geq 0.5$	$2 > \beta \geq 1$	$3 > \beta \geq 2$	$5 > \beta \geq 3$
Poor	$\beta < 0.5$	$\beta < 1$	$\beta < 2$	$\beta < 3$

Assay Guidance Manual

Assay Guidance Manual

Table 3: Interpretation of AUC values

Value	Interpretation
AUC = 1	A "perfect" assay.
AUC = 0.5	Poor: Performance is only as good as random chance.
AUC < 0.5	Worse than random chance.